Spring 2018

A Review of the Utility of Bayesian Network Models

Luke Magyar
lrm58@zips.uakron.edu

A Review of the Utility of Bayesian Network Models

Luke Magyar

Department of Mathematics

Honors Research Project

Submitted to

The Honors College

Approved:	Accepted:
_____ Date _____ Honors Project Sponsor (signed)	_____ Date _____ Department Head (signed)
_____ Honors Project Sponsor (printed)	_____ Department Head (printed)
_____ Date _____ Reader (signed)	_____ Date _____ Honors Faculty Advisor (signed)
_____ Reader (printed)	_____ Honors Faculty Advisor (printed)
_____ Date _____ Reader (signed)	_____ Date _____ Dean, Honors College
_____ Reader (printed)	

Abstract

Bayesian Networks are probabilistic models built from conditional probability tables that relate two observable instances to one another in parent-child fashion. The networks' strength lies in their ability to use inferential logic to make likelihood assessments about a parent node based on an observation of its child. Additionally, they make it very easy to combine quantitative data with qualitative knowledge from industry experts. These abilities make them very attractive for use as formulation tools in the paint and rubber industries. Paint and rubber formulation has long proven to be a challenging task because companies have a difficult time compiling the data from all their formulators- data that often contains large amounts of opinion. This paper seeks to define Bayesian Networks and a few inferential operations using them, and then to apply these methods to three distinct industry problems. This paper explores applications including: (1) marketing, (2) expert knowledge collection, and (3) a traditional formulation study. This paper is submitted as part of graduation requirements for the University of Akron Williams Honors College, 2018.

Acknowledgements

Particular thanks are due to the team of advisors who directed this project including Dr. Nao Mimoto, Dr. Curtis Clemons, and Dr. Jerry Young. Many hours were spent by them to progress this project. Further guidance and expert recommendation was given by Dr. Kevin Kreider and Dr. Scott Lillard. In addition, the work by Kaylee Sutton and her team at Sherwin-Williams is particularly appreciated for taking the time to provide much of the data and inspiration for this project. Lastly, a great amount of support was given by Dr. Qixin Zhou and her team of student researchers in providing the data and analytical recommendations for the epoxy study.

I thank you all very much.

Introduction and Theory

A Bayesian network is a statistical model, used for decision making, built from sets of conditionally dependent variables represented as nodes related to one another in "parent-child" fashion. These relationships are represented with a probability function, most commonly the conditional probability function. Developing the network in this manner allows it to answer questions of the form "If X event occurs, what is the probability that Y [dependent on X] will also occur." Developing these relationships requires the use of data or expert opinion to build a conditional probability table (CPT), wherein a probability for each output (possible state of a node) is calculated as a function of the states of its parents. This data can take the form of empirical data or computer generated simulations as well as organized expert opinion [8]. The use of expert opinion makes Bayesian networks particularly useful in situations where data can be expensive or difficult to acquire or in cases where data results are largely qualitative. **Figure 1** gives a simple example to illustrate the CPT for each node of a small system.

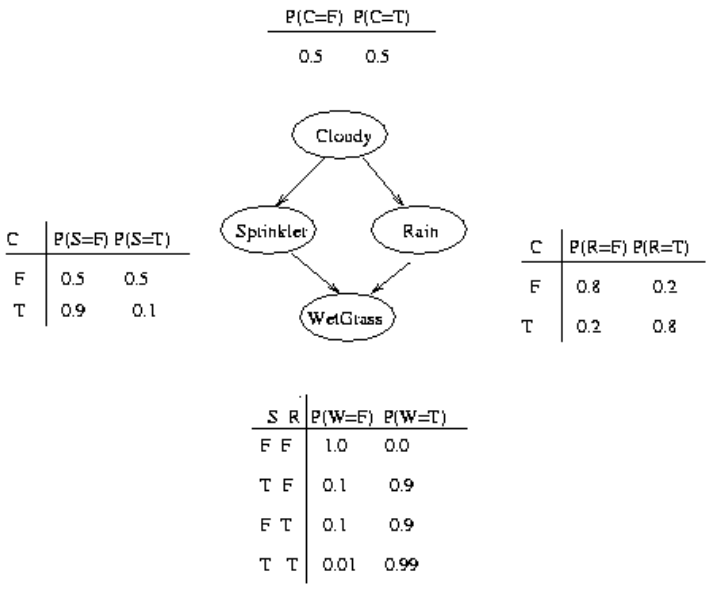


Figure 1
A sample Bayes Net showing the conditional probability tables for each node. [8]

In this system, the child node attempts to answer the True/False question "Is the grass wet?" To calculate this probabilistically, the model takes other T/F questions into account as parent nodes. These are "Is the sprinkler on?" and "Is it raining?" which are in turn the child nodes of the initial parent "Is it cloudy?" Hypothetical "data" has been input into the CPT's for illustration: in the

Cloudy node, it is seen that there is a 50% chance of the answer being either *True* or *False*. Next the *Sprinkler* and *Rain* nodes form probability tables without knowing the state of *Cloudy*. For the *Rain* node, data indicates that if it is cloudy (*True*) then there is an 80% chance of rain. If *Cloudy* is *False* then there is only a 20% chance of *Rain* being *True*. Similar data is added for the *Sprinkler* node, and the combination of these two nodes are used for *Wet Grass*; here each parent has 2 possible states (*True* or *False*) as does the child node. This means 8 [2x2x2] pieces of data are needed to fill out the table with every possible outcome.

In this format, we can only determine the answer to the question "Is the grass wet?" if we observe the state of the other three nodes. Often we would like to view each of the nodes as probabilities and observe the overall probability of the child node. In our example this allows us to determine the overall probability of each node being true based on the initial known of a 50% chance of it being cloudy. **Table 1** demonstrates how to use the values in the CPT's to calculate probabilities for *Cloudy*'s two child nodes.

Cloudy?		Sprinkler?		Total	Rain?		Total
Value	Prob	Value	Prob	Likelihood	Value	Prob	Likelihood
F	0.5	T	0.5	0.25	T	0.2	0.1
T	0.5	T	0.1	0.05	T	0.8	0.4
			P(S=T)	0.3		P(R=T)	0.5
F	0.5	F	0.5	0.25	F	0.8	0.4
T	0.5	F	0.9	0.45	F	0.2	0.1
			P(S=F)	0.7		P(R=F)	0.5

Table 1

Table calculating the likelihood of the *Sprinkler* and *Rain* nodes having the value *True*. Note that the values in the "Prob" columns are CPT entries corresponding to the T/F state in the Value columns (e.g. the "Prob" value in the first row for the *Sprinkler?* column can be found in the CPT in **Figure 1** where C=F and S=T). The "Total Likelihood" column lists the overall likelihood of each state occurring. These values are arrived at by multiplying the *Cloudy* probability with that of either *Sprinkler* or *Rain* for each row. The overall likelihood of *Sprinkler* or *Rain* having the value of *True* is the sum of the likelihoods for each case where the node has a value of *True*. These are noted as subtotals in the colored cells of the figure.

Based on the analysis in **Table 1** a 50/50 chance of cloudiness indicates a 30% chance of the sprinkler being on and a 50% chance of it raining. This same method of analysis can then be applied to the child node *Wet Grass* as seen in **Table 2** where an overall chance of the grass being wet is calculated as 64.7%.

Cloudy?		Sprinkler?		Rain?		Wet?		Total
Value	Prob	Value	Prob	Value	Prob	Value	Prob	Likelihood
F	0.5	F	0.5	F	0.8	T	0	0
F	0.5	T	0.5	F	0.8	T	0.9	0.18
F	0.5	F	0.5	T	0.2	T	0.9	0.045
F	0.5	T	0.5	T	0.2	T	0.99	0.0495
T	0.5	F	0.9	F	0.2	T	0	0
T	0.5	T	0.1	F	0.2	T	0.9	0.009
T	0.5	F	0.9	T	0.8	T	0.9	0.324
T	0.5	T	0.1	T	0.8	T	0.99	0.0396
							P(W=T)	0.6471
F	0.5	F	0.5	F	0.8	F	1	0.2
F	0.5	T	0.5	F	0.8	F	0.1	0.02
F	0.5	F	0.5	T	0.2	F	0.1	0.005
F	0.5	T	0.5	T	0.2	F	0.01	0.0005
T	0.5	F	0.9	F	0.2	F	1	0.09
T	0.5	T	0.1	F	0.2	F	0.1	0.001
T	0.5	F	0.9	T	0.8	F	0.1	0.036
T	0.5	T	0.1	T	0.8	F	0.01	0.0004
							P(W=F)	0.3529

Table 2

Table calculating the likelihood of *Wet Grass* having a value of *True* or *False*. Because this node has two parents (which in turn share a parent) this analysis requires testing 16 cases. The likelihood of each case occurring, listed in the "Total Likelihood" column, is the product of the four probabilities in the row. These are summed for all cases where *Wet Grass* is *True* to provide the subtotal labeled P(W=T). Similarly, the likelihood of the node being false is calculated in P(W=F).

The size of the table required for **Table 2** illustrates a shortcoming of Bayesian Nets in that large amounts of data and computations are required. To reduce the number of hand calculations, computer software can be used. This paper makes use of the commercially available software Netica from Norsys [9]. **Figure 2** shows this same system from **Figure 1** built using Netica; here we use the initial input of a 50/50% chance of it being cloudy, which yields a 50/50% chance of it raining, but only a 30% chance of the sprinkler being on. Cumulatively this gives a 64.7% chance of the grass being wet. Here it is seen that Netica is able to replicate the results from **Tables 1** and **2**.

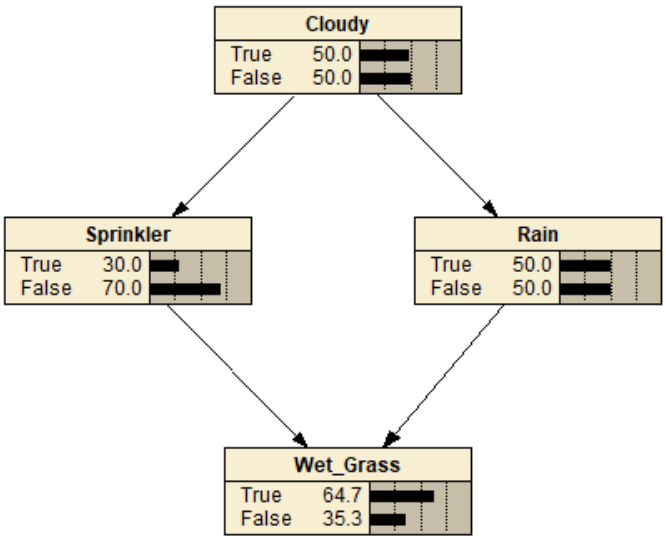


Figure 2
The system from **Figure 1** transposed to the Netica Program. Note the auto-calculated probabilities in the child nodes and compare them with **Tables 1** and **2**.

In **Figure 3** it is demonstrated how changing the initial probability of the *Cloudy* node affects the rest of the network. Here the higher likelihood of cloudiness has reduced the likelihood of the sprinkler being used but raised the likelihoods for both the *Rain* and *Wet Grass* nodes. The ability to quickly test many potential states make Netica particularly useful for analyzing a system.

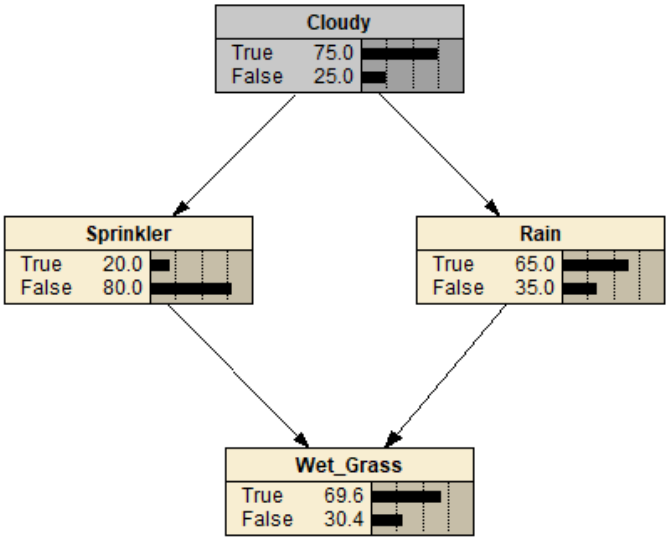


Figure 3
Netica network with adjusted probabilities in the *Cloudy* node.

Now we can use the model from **Figure 2** to make choices about what we know to be true, such as in **Figure 4** which shows two different known states that have affected how the other nodes are calculated. Note when we tell the model that it is not raining, the probability of it being cloudy changed to reflect this. This is referred to as "Explaining Away" and it indicates that the two nodes are not fully independent from one another because of their shared child [8]. Another important application of Bayesian Networks is the use of "Inferencing" where a known value of a child node is used to calculate the value of the parents [5]. **Figure 5** demonstrates this by picking a case where we observe that the grass is wet. The model then calculates that there is a 43% chance that the sprinkler was on, and a 70.8% chance that it had rained. This can be thought of as the model inferring the cause of an observed state from the values in the CPT. Here this means that when we see the grass is wet and make no other observations about the weather, we can say that there is a higher likelihood that it was caused by rain than by the sprinkler. In turn, there is also a higher likelihood of it being cloudy.

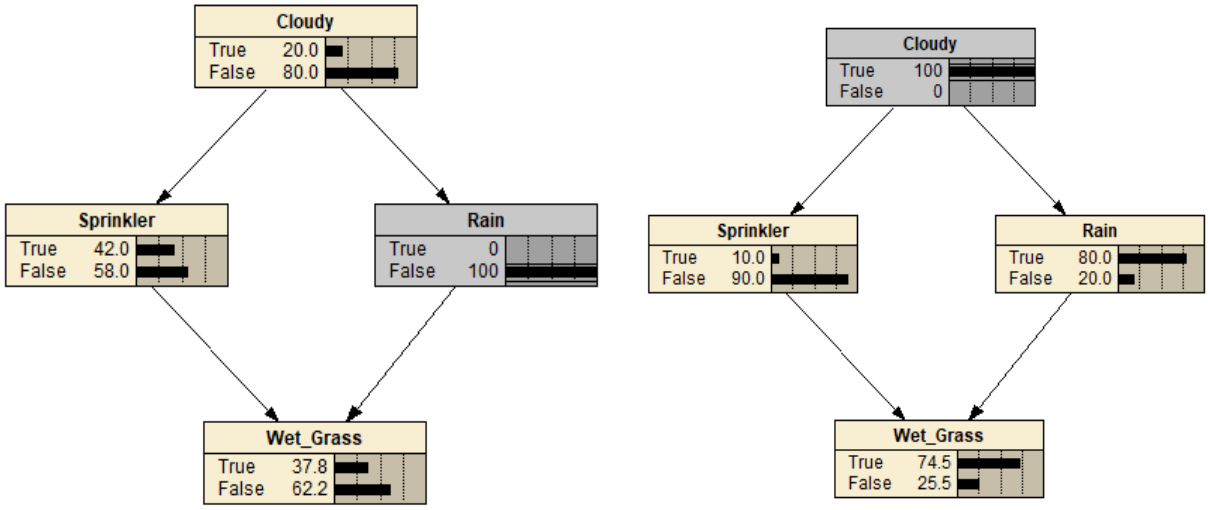


Figure 4

A pair of screen-captures from Netica, using the initial model in **Figure 2.** On the left, an observation that it is not raining has been made. On the right, an observation that it is cloudy has been made. The effects on the probability of the other statements being true is updated in each node.

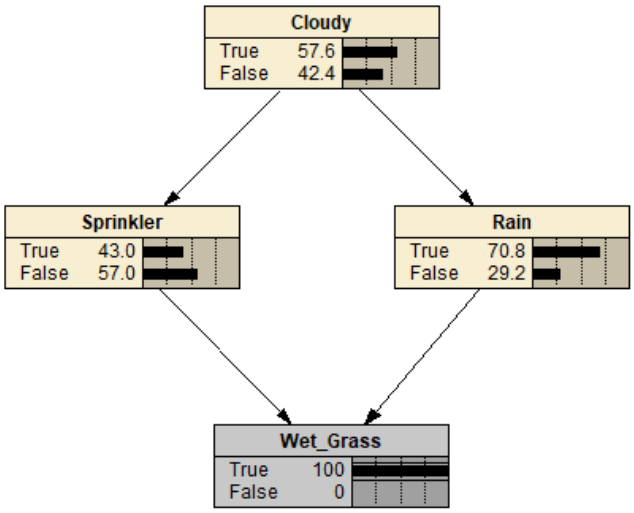


Figure 5

An example of using Bayesian Networks for inferencing. Here it has been observed that the grass is wet and the model uses the CPT's to infer the values of the other nodes.

Bayesian Networks have had a wide variety of novel research and actual industrial implementation, especially since the late 1990s after widespread computing made collecting large data sets realistic [8]. A few of the more recent findings in this field are presented here to gain an insight into the work being done with Bayesian Networks. In 2014, Ayello, Jain, Sridhar, and Koch applied probabilistic modeling to the problem of corrosion. In this paper, the authors developed a model to assess the corrosion potential of pipelines based on many causal factors [1]. They sought to use their model's inferential abilities to determine root causes of known corrosion failures, as well as to identify sites with high likelihoods of failure. In 2016, Fu and Deng, et.al. published a paper investigating the use of Bayesian Networks to identify biological pathways [4]. They built a model using interaction data for proteins and genetic structures and then used it to predict signaling pathways. The initial indications of this study were that Bayesian Networks were more successful at this task than traditional methods and at a greatly reduced difficulty. Dal Ferro, Quinn, and Morari published a paper in 2018 in which they used a Bayesian Network to model the dynamics of soil organic carbon (SOC) in agricultural land [2]. Their model was able to replicate the results of real-world data for both SOC accumulation and depletion, as well as to develop management scenarios for preserving desired levels. Also in 2018, Xie and Gao, et.al. used Bayesian Networks to develop predictive models for autonomous vehicles [12]. Their networks were used to model lane-changing maneuvers of driverless cars by utilizing physics based prediction as well as inferential Bayes methods for random scenarios. The tools demonstrated in this example as well as the thought processes shown in these papers have been applied to three industrially significant problems in an effort to demonstrate the usefulness of this type of analysis. These problems include: A Sherwin-Williams marketing tool, a Sherwin-Williams expert opinion study, and an epoxy formulation study.

Sherwin-Williams Marketing Tool

This model utilizes product rating data from a placard, seen in **Figure 6**, displayed inside the Sherwin-Williams store located in Cuyahoga Falls, Ohio [11]. This data was used to build CPTs and develop a Bayesian Network of this system with customer choice in mind. This system was designed in Netica, as seen in **Figure 7**, as a network having eight nodes: one for the paint, one for each of the properties rated in **Figure 6** as well as the cost, and a final node used to sort the paints into thirds based on what the user finds important. For each property node, the data was condensed from a five star system to four bins, where A=5-stars, B=4.5-stars, C=4-stars, and D=3.5-stars. This was done to minimize the bias of having no paints with less than 3.5-star ratings in any category. **Table 3** demonstrates the CPT for the *Mildew & Moisture Resistance* node, where each paint was taken to have a 100% chance of existing in the bin described by the star rating in **Figure 6.**

Figure 6

A placard displayed in a Sherwin-Williams consumer store in Cuyahoga Falls. This is used as the source of the data used to develop the model in this section.

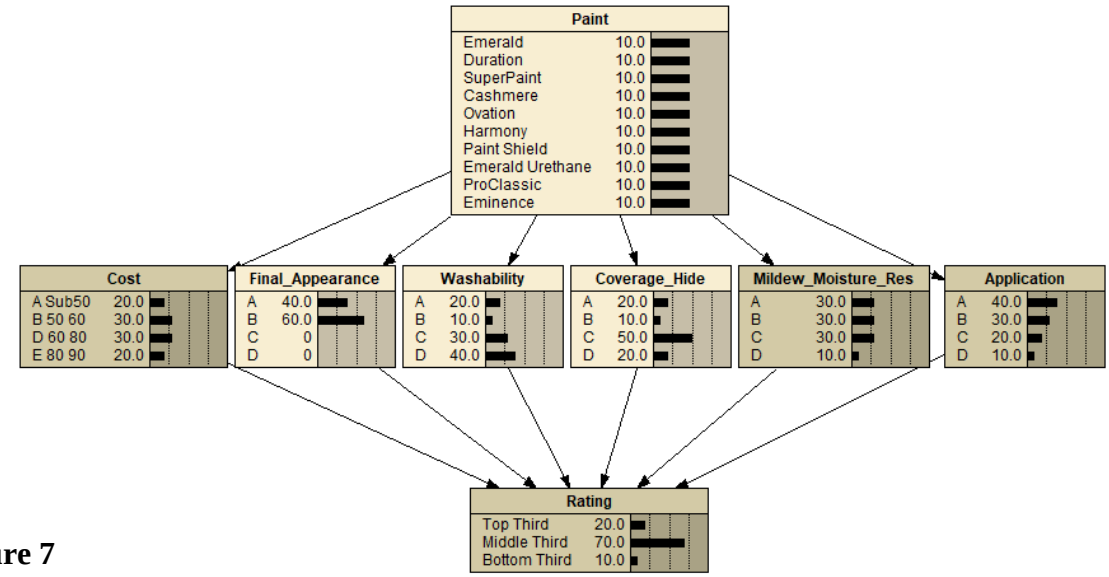


Figure 7

Netica model prepared from the data in **Figure 6** where the properties have been converted from a 5-star rating system to four discrete bins. An *A* rating corresponds to "Best", and *D* to "Worst". The CPT for the *Rating* node is generated by an excel table, using a weighting system decided by the user. The *Top Third* rating indicates the "best choice" paints for the users preferences.

Mildew & Moisture Resistance				
Paint	A	B	C	D
Emerald	1	0	0	0
Duration	1	0	0	0
SuperPaint	0	0	1	0
Cashmere	0	0	1	0
Ovation	0	0	1	0
Harmony	0	1	0	0
Paint Shield	0	1	0	0
Emerald Urethane	1	0	0	0
ProClassic	0	1	0	0
Eminence	0	0	0	1

Table 3

CPT for the *Mildew & Moisture Resistance* node from **Figure 7.** The data comes from **Figure 6** where *Emerald*, *Duration*, and *Emerald Urethane* each received 5-star ratings, and as such are listed as the best "A" rating here. *Eminence* is listed in the "D" column with a 3.5-star rating.

The CPT for the rating node is built using an excel table and user input multipliers for each property. This table has 4098 rows, one for each possible state of the parent nodes [4^6 = 4098]. These rows are then scored based on the sum of each "good" result in each node; for example, the "best" possible coating has an A rating in each category, and is cheaper than the others (note that this coating does not exist). These scores are then weighted by the user ratings of which property

is most important. Finally, these scores are used to create the CPT by dividing each of the 4098 states into an upper, middle, and lower third. **Table 4** shows part of this spreadsheet and gives addition details about the methodology. **Figures 8** and **9** show the effect of using different rating schemes in **Table 4**. The model in **Figure 8** equally rates each property, showing that two paints, *Emerald* and *Emerald Urethane* are contained in the upper third of all paints. **Figure 9** demonstrates how using a weighting scheme that heavily favors cheaper paints pushes the two expensive paints *Emerald* and *Emerald Urethane* into the middle third despite their superior performance in other categories. This corresponds the rating scheme demonstrated in **Table 4**. No paints make it into the upper third in this weighting scheme.

Cost	Appearance	Washability	Coverage	Mildew	Application
5	1	1	1	1	1

Cost	Appearance	Washability	Coverage	Mildew	Application	Score	1st Third	2nd Third	3rd Third
						30	30	20	9
4	4	4	4	4	4	40	100	0	0
4	4	4	4	4	3	39	100	0	0
4	4	4	4	4	2	38	100	0	0
4	4	4	4	4	1	37	100	0	0
4	4	4	4	3	4	39	100	0	0
4	4	4	4	3	3	38	100	0	0
4	4	4	4	3	2	37	100	0	0
4	4	4	4	3	1	36	100	0	0
4	4	4	4	2	4	38	100	0	0
4	4	4	4	2	3	37	100	0	0
4	4	4	4	2	2	36	100	0	0
4	4	4	4	2	1	35	100	0	0
4	4	4	4	1	4	37	100	0	0
4	4	4	4	1	3	36	100	0	0
4	4	4	4	1	2	35	100	0	0
4	4	4	4	1	1	34	100	0	0
4	4	4	3	4	4	39	100	0	0
4	4	4	3	4	3	38	100	0	0
4	4	4	3	4	2	37	100	0	0
4	4	4	3	4	1	36	100	0	0

Table 4

This is the rating table which has been truncated due to the length of the full table. The box at the top shows the user selected ratings; here, low cost has been heavily weighted as the desirable component. Note that any numbers can be used for this rating system, including zero which corresponds to no preference. The lower box makes up the actual CPT, where the colored numbers to the left indicate the different states that are the rows of this CPT. Here, the ABCD notation has been converted to numbers with 4 being the best and 1 the worst. The score column gives the score for each case by multiplying the property values by their multipliers and summing them. The yellow highlighted cell is the range of the scores and is used to develop the thresholds seen next to it for the three thirds. Each case is then sorted into one of the three thirds by comparing its score with the three thresholds. Here the 100's seen in the "1st Third" column indicate that a paint has a 100% probability of being in that third if it falls onto that row of the CPT.

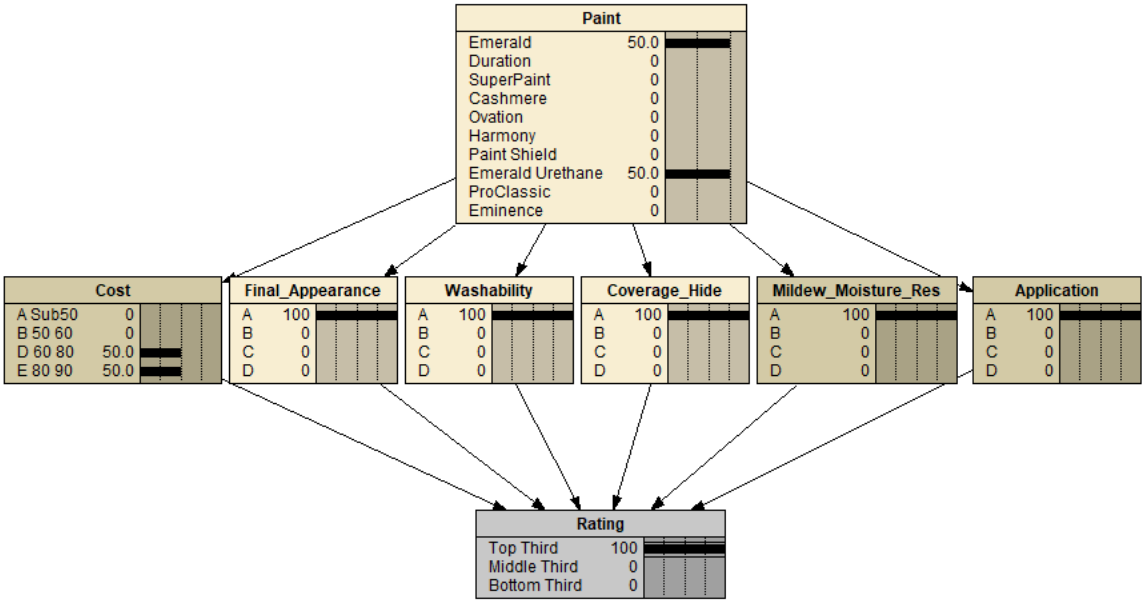


Figure 8

A Netica model that demonstrates user input that equally weights all paint properties. The *Top Third* rating is selected in this case to reveal the paints in this subgroup.

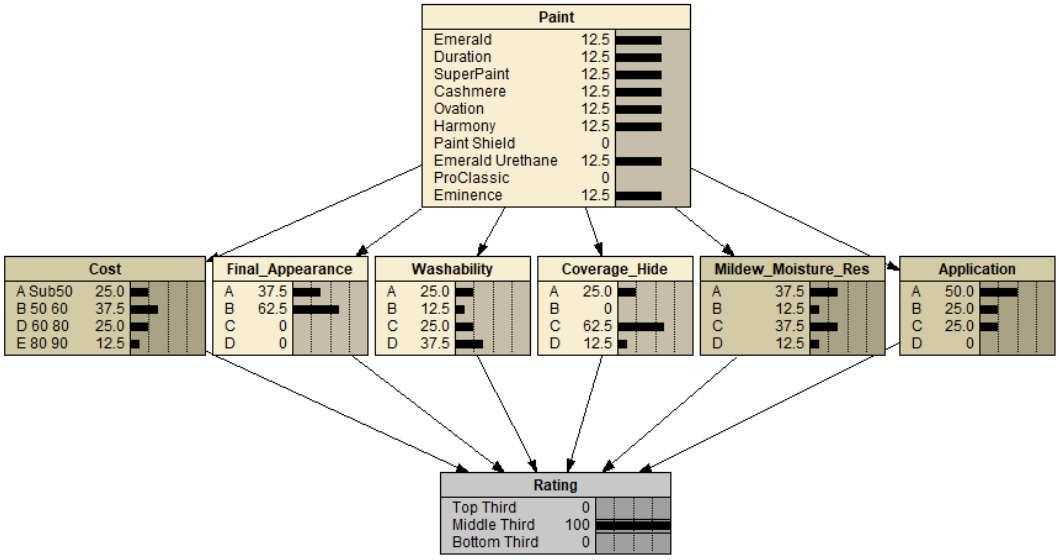


Figure 9

The Netica model that demonstrates user input that heavily favors low cost paints. The *Middle Third* rating is selected in this case to reveal the paints in this subgroup. No paints achieved a *Top Third* rating with this weighting scheme.

This method for setting up the Bayesian Net demonstrates how inferencing can be used as a design feature. In this model, the user rates the properties it wants and the model infers which paints would produce them. Even in its current rudimentary state, this model is potentially easier for the customer to use than the placard in **Figure 6**. Here they can use the rating system to visually eliminate properties that they do not care about, as well as be told what their best options are. This tool would make it easier for companies to market products with complicated decision making processes to their customers.

Sherwin-Williams Pigment Tool

Another model built for this project utilized empirical observations about how pigments affected various paint properties. This data was compiled and provided by Kaylee Sutton from Sherwin-Williams and represents a review of industry-expert knowledge and observation [10]. This model illustrates the strength of Bayesian Networks, because abstract data that utilizes an expert's beliefs can be used alongside scientific data. This is valuable for applications such as paint or polymer formulation, where many overlapping variables can affect desired properties making it difficult to produce factorial experimental designs. **Table 5** is used as the raw data for this model, taken as expert opinion compiled by Kaylee Sutton using knowledge taken internally from Sherwin-Williams, as well as from commonly accepted sources such as Joseph V. Koleske's *Paint and Coating Testing Manual* [6]. **Figure 10** shows the model created from this data. The CPTs for this model were developed as in **Table 6**. This model illustrates the use of Bayesian nets as an alternative method of viewing tabular data in a more illustrative way. Additional data showing how differing amounts of each pigment component affects properties would make this model more extensive.

	Calcium Carbonate 5µm	Hydrous Kaolin 0.5µm	Hydrous Kaolin 4.8µm	Calcined Kaolin	Talc 8µm	Silica 5+10µm	Feldspar	Nepheline Syentite
Hiding	Very Low	Very High	Moderate	Very High	Moderate	Moderate	Low	Low
Gloss	Low	Moderate	Low	Low	Low	Low	Low	Low
Chemical Resistance	Very poor	Very Good	Very Good	Good	Good	Excellent	Excellent	Excellent
Abrasion Resistance	Fair	Poor	Poor	Excellent	Good	Good	Very Good	Very Good
Viscosity	Very Low	High	Low	High	Moderate	Low	Low	Low
Stability	Poor to Good	Excellent	Excellent	Very Good	Poor to Good	Excellent	Excellent	Excellent
Ease of Brushing	Fair	Excellent	Excellent	Good	Fair	Good	Good	Good
Leveling	Excellent	Very Good	Very Good	Good	Fair	Very Good	Very Good	Very Good

Table 5

Data provided by Sherwin-Williams that represents expert opinions about the abilities of each pigment to affect the properties in the left-hand column. These values were used to fill the CPTs in **Figure 10** as seen in **Table 6**.

Hiding					
Component	Very Low	Low	Moderate	High	Very High
Calcium Carbonate	1	0	0	0	0
Hydrous Kaolin Low	0	0	0	0	1
Hydrous Kaolin High	0	0	1	0	0
Calcined Kaolin	0	0	0	0	1
Talc	0	0	1	0	0
Silica	0	0	1	0	0
Feldspar	0	1	0	0	0
Nepheline Syenite	0	1	0	0	0

Table 6
Sample CPT for the *Hiding* node for **Figure 10**.

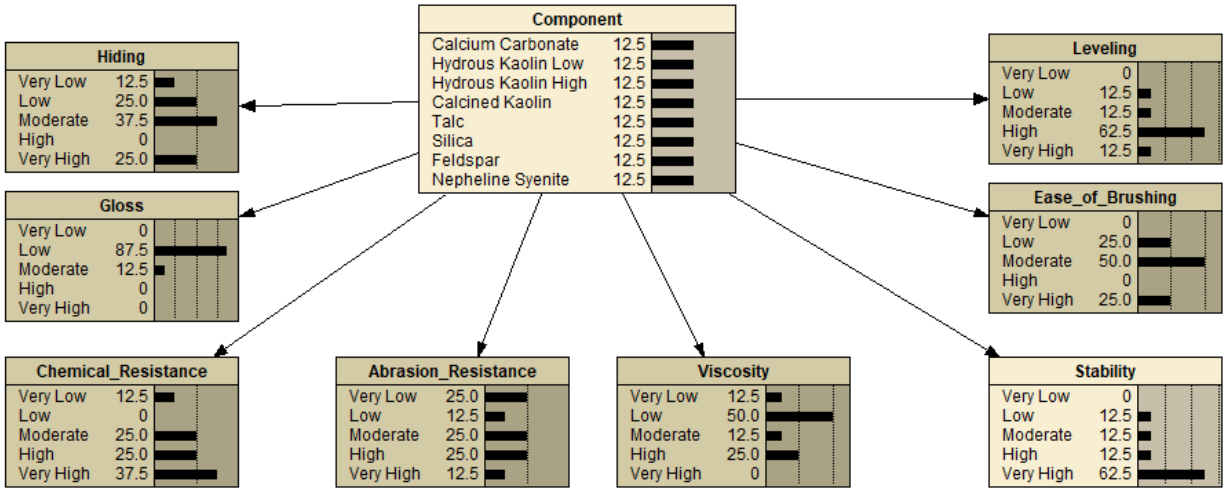


Figure 10
A Netica model displaying the performance of a variety paint pigments and fillers. This model represents a distillation of expert knowledge into a condensed model.

In this Netica model, selecting a known in the *Component* node will simply reproduce the results in **Table 5**. The true strength of this network though is demonstrated in **Figure 11** where we have told the model that we equally desire either a low or a very low viscosity. The model then infers which paints to use, as well as the likelihoods we can expect for the other properties.

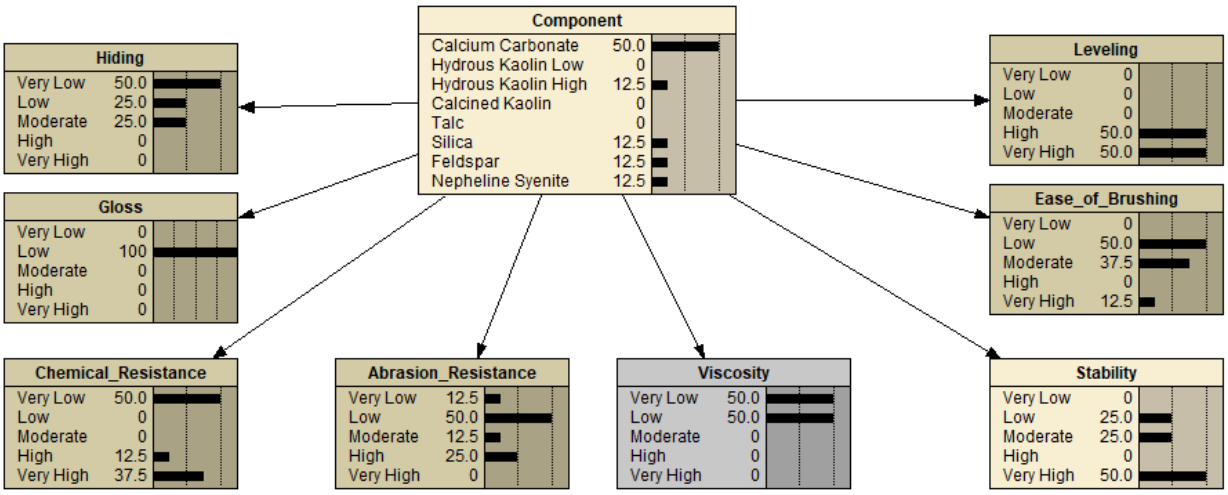


Figure 11

Expert opinion model used to show how a desired viscosity range affects how the other nodes. Here Calcium Carbonate has been inferred as the best choice, and the rest of the property nodes have been recalculated based on the probabilities in the *Component* node.

This style of analysis shows how Bayesian Networks can replace tabular data into a dynamic model to aid decision making. Additionally, it is able to condense industry knowledge without quantitative data into a useable form. Combining this model with a formulation study would increase the accuracy of the model by providing probabilities for the CPTs (e.g. testing could show that paints with *Talc* as a pigment may have a 90% chance of being rated "Good" for *Abrasion Resistance* and a 10% chance of being rated "Fair"). Bayesian Networks allow for both types of data to be easily mixed so that expensive and time consuming testing need only be performed where it is deemed most important; the rest of the network can rely on expert opinion.

Epoxy Formulation Study

Lastly, this modeling study utilized data from a coatings formulation study performed by Dr. Qixin Zhou's research group to develop a predictive Bayesian Network [8]. The formulation study sought to categorize the effect of changing various formula components of corrosion preventative epoxy coatings. These parameters include the choice of solvent, solid concentration, and pigment concentration. To categorize the effects of changing these parameters, three tests were performed for each coating: (1) adhesion to a metal substrate, (2) color change (ΔE*) under UV exposure, and (3) electrical impedance spectroscopy (EIS), which indicates the ease at which water leaches into the coating. Due to the nature of some of the coatings, data is not available in all cases; for these instances, an "Unknown" designation is used. **Table 6** shows the testing matrix for this study and explains the lack of data for certain cases.

	Pigment Concentration (%)	Solid Content (%)	Solvent	Notes	Adhesion	ΔE	EIS	Netica Key
Tier 1 (Vary Solvent)	0	75	Acetone	Good	Yes	Yes	Yes	Acetone_0_75
	0	75	Ethanol	Not suitable as epoxy binder	No	No	No	Ethanol_0_75
	0	75	Xylene	Good	No	Yes	No	Xylene_0_75
Tier 2 (Vary Solids)	2	80	Acetone	High viscosity	No	No	No	Acetone_2_80
	2	75	Acetone	Good, easy to apply	Yes	Yes	Yes	Acetone_2_75
	2	70	Acetone	Good, low viscosity	No	No	No	Acetone_2_70
Tier 3 (Vary Pigment)	1	75	Acetone	Easy to disperse	Yes	Yes	Yes	Acetone_1_75
	2	75	Acetone	Easy to disperse	Yes	Yes	Yes	Acetone_2_75
	4	75	Acetone	Not easy to disperse	Yes	Yes	Yes	Acetone_4_75

Table 6

Testing matrix for this formulation study. In Tier 1, a constant pigment concentration and solids content was used while varying the solvent. In this case, ethanol was deemed unsuitable and no testing was possible due to the poor nature of the coatings. Xylene was also eliminated because of its toxicity after it showed similar ΔE* and application results to the acetone coating, indicating that it had no further benefit to outweigh the dangers of its use [13]. Similarly in Tier 2, the 80% and 70% solid content coatings were not tested due to the superior application ability of the 75% coating. The *Netica Key* column shows the names assigned to each coating in the Netica model. Note that *Acetone_2_75* appears in both Tier 2 and Tier 3.

The average values from the Adhesion test are displayed in **Table 7**. Three trials were performed for each coating to achieve these numbers. None of the tested coatings had issues with adhesion, though it should be noted that many of the untested coatings, particularly the ethanol solvated ones, would have performed poorly. For this reason, a Good/Poor rating was not appropriate so Best/Okay was used instead. From these results, medium amounts of pigment appear to produce better adhesion, although the total difference may be negligible. **Table 8** displays the measured color change for each coating after 20-days exposure in a QUV chamber which accelerates UV aging of coatings. The reported value is the Euclidean "distance" between the starting color point and the ending color point in a three dimensional color space. Literature suggests that the minimum perceptible ΔE value is 2.3 [7]. Any coating below 3.0 after 20 days was said to be "Good". From this data it is clear that increasing pigment concentration improves the color retention. The two coatings with zero pigment but different solvents, acetone and xylene respectively, suggest that acetone is better than xylene at resisting color change.

Coating	Average Pull-off Pressure (psi)	Rating
Acetone_0_75	235	Okay
Acetone_1_75	255	Best
Acetone_2_75	243	Best
Acetone_4_75	237	Okay

Table 7
Average values for adhesion strength for the four coatings tested for this property. All four were acceptable, so the two highest values were given the *Best* designation, and the others were given *Okay*.

Coating	20 Day ΔE Value	Rating
Acetone_0_75	21.83	Poor
Xylene_0_75	25.87	Poor
Acetone_1_75	8.41	Poor
Acetone_2_75	2.59	Good
Acetone_4_75	1.98	Good

Table 8
Ultimate color change values for the five tested coatings. Reported ΔE values are the International Commission on Illumination (CIELAB) standard which measures the "distance" between two colors in a Euclidean space of colors. The critical value for rating the coatings as good or poor was taken to be 3.0; this is above the threshold perceptible to the human eye [7].

In order to categorize the results of the EIS tests, some analysis had to be performed. Traditionally, EIS test results are used to rank coatings by which has the highest impedance (Zmod) value at a critical frequency, typically the lowest tested frequency. For the purposes of this study, however, it was necessary to also attempt to categorize how quickly the impedance degraded over time, which indicates that water has leached into the coating, thus compromising the corrosion inhibitance [13]. **Figure 12** demonstrates the need for this question because these traditional EIS plots make it difficult to pick the better coating. In this figure, the two coatings have a different Zmod value after the first test day, however the Zmod values after longer times are very similar. Based on this, it can be difficult to say that one is better than the other and therefore the change over time should be factored in to the analysis. Further complicating the matter is a natural variance in EIS testing between identical coatings. EIS testing is highly sensitive to factors such as ambient electrical interference from motors and computers, minor variation in cell lead placement, and defects or inconsistencies in the coating surface [3]. A way to categorize the rate at which a coating changed with respect to time could help to mitigate the uncertainty created by this potential variance in the initial Zmod value.

To categorize the change over time, a gradient method was used to calculate the change between two test days at each frequency. This change was divided by the time change between test-points as well as by the initial Zmod value. This had the effect of normalizing the data sets to help compare two coatings with different initial Zmod values. **Figure 13** shows the same test data as **Figure 12** but in this normalized gradient form. Combining the information from both graphs, it can be seen that the coating with 2% pigment concentration had a higher initial Zmod value, but also degraded faster than the coating with 0% pigment. These competing properties make it necessary for a system to combine them based on the formulator's desire. **Table 9** demonstrates one such potential system where the coatings were ranked from 4 (best) to 1 (worst) in both initial Zmod and initial gradient. Then they were combined into an average value for each coating by weighting the Zmod by 90% and the gradient by 10%. This produced a range of scores, from which the bottom two were taken as "Poor" and the top were "Good". This ranking system is arbitrary and represents a "best guess" first attempt at developing a model for EIS. A degree of sensitivity is present in this ranking system such that if the weighting system is shifted to 60% initial, 40% gradient then the 1% pigment coating is said to outperform the 2% coating. These multiplier values

should be refined by collecting more data and adjusting the Good/Poor ratings to agree with expert opinion and other testing methods. The decisions made for EIS data are summarized in **Table 9**.

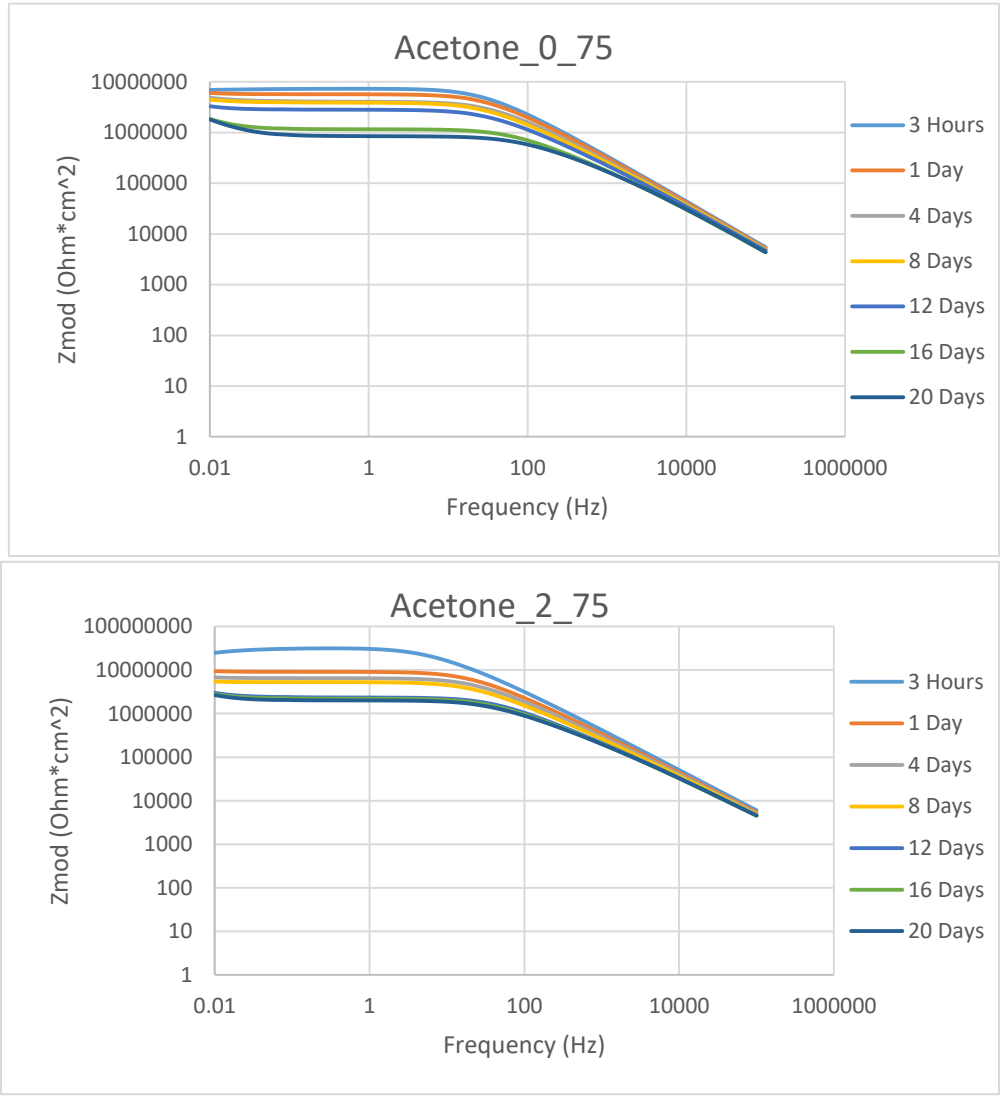
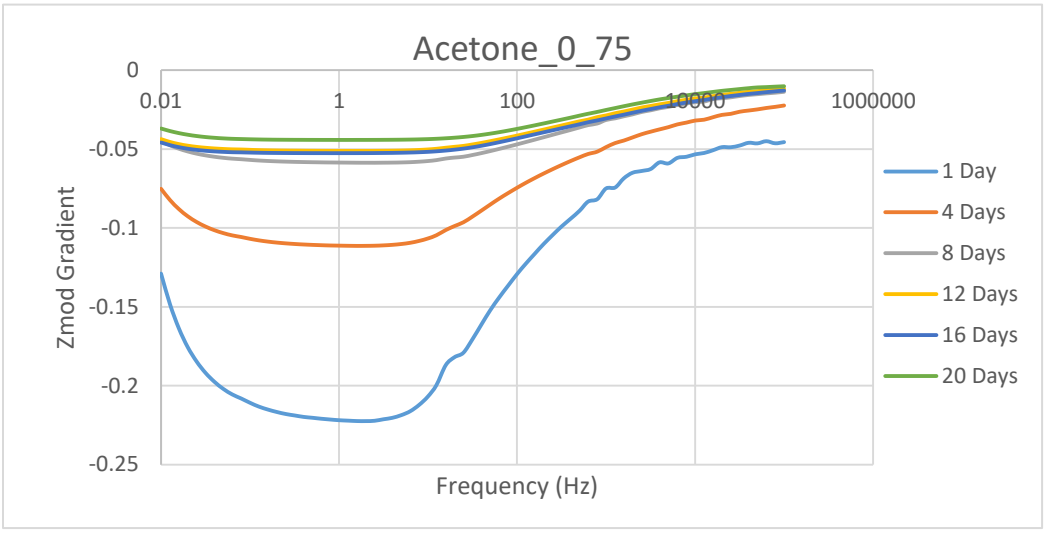


Figure 12

EIS plots for two of the coatings in this study, both using Acetone and 75% solid content. The top plot indicates 0% pigment conc. and the bottom corresponds to 2%. Note the log-log axes, and that the lower graph starts at higher initial value of Zmod. Traditionally, higher initial values of Zmod have been used to indicate a coating with lower water permeability properties.



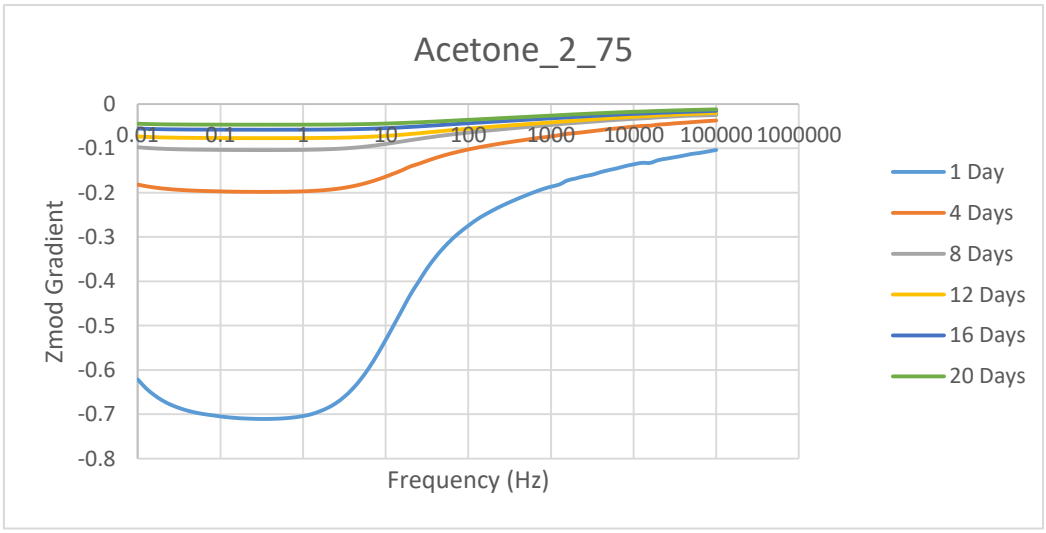


Figure 13

Plots used for comparison of the impedance breakdown of two coatings (corresponding to an increase in water uptake). Note the horizontal axis is log scale. The vertical axis displays the calculated, normalized change in Zmod at each frequency over time. By this comparison, the lower graph is said to have performed worse because of the large initial change at Day 1.

Coating	Initial Value	Score	Peak Gradient	Score	Overall Score	Rating
Acetone_0_75	6.92E+06	1	-0.2225	3	1.2	Poor
Acetone_1_75	8.32E+06	2	-0.1665	4	2.2	Poor
Acetone_2_75	2.49E+07	3	-0.7105	2	2.9	Good
Acetone_4_75	4.38E+07	4	-0.8780	1	3.7	Good

Table 9

Summary of EIS data for four coatings. The initial value column lists the day zero (measured 3 hours after application) Zmod at the initial frequency of 0.01 Hz. The peak gradient measures the largest first-day gradient seen at any frequency for a coating. These are scored such that the best coating has 4 points. The overall score is adjusted by the arbitrary choice of 90% Initial, 10% Gradient. The two highest scoring coatings received a "Good" rating.

All three properties were used to create the CPT tables that are used to develop **Figure 14**. The middle tier of nodes have CPT tables that point to the *Formulas* node, they are included because not every possible combination of solvent, pigment concentration, and solid content is included in the *Formulas* node. This method allows for additional coatings to be added to this study without changing the structure of lower CPTs. Selecting one the coatings in the *Formulas* node will produce its exact formulation in the three child nodes. The bottom nodes have CPT tables based on every possible formula (3 solvents x 3 pigment conc. x 3 solids content = 27 formulas). For this reason, a large number of these states are unknown, as data has been collected only for the few coatings described in **Table 6**. **Figure 15** shows that the model picks the coating using Acetone, 2% pigment, and 75% solids as the ideal coating. This model is meant as a first attempt at developing a Bayesian Network for a formulation study, and the intent is for it to expand as more data becomes available from Dr. Zhou's study. In turn, the results of this analysis can be used to direct the formulation study for future development.

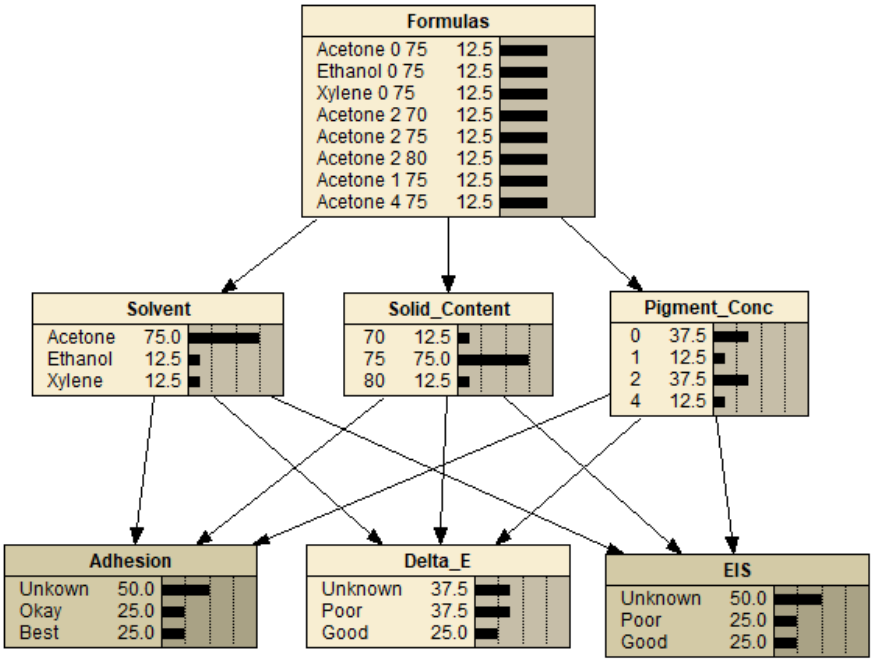


Figure 14

Netica model of the epoxy formulation data.

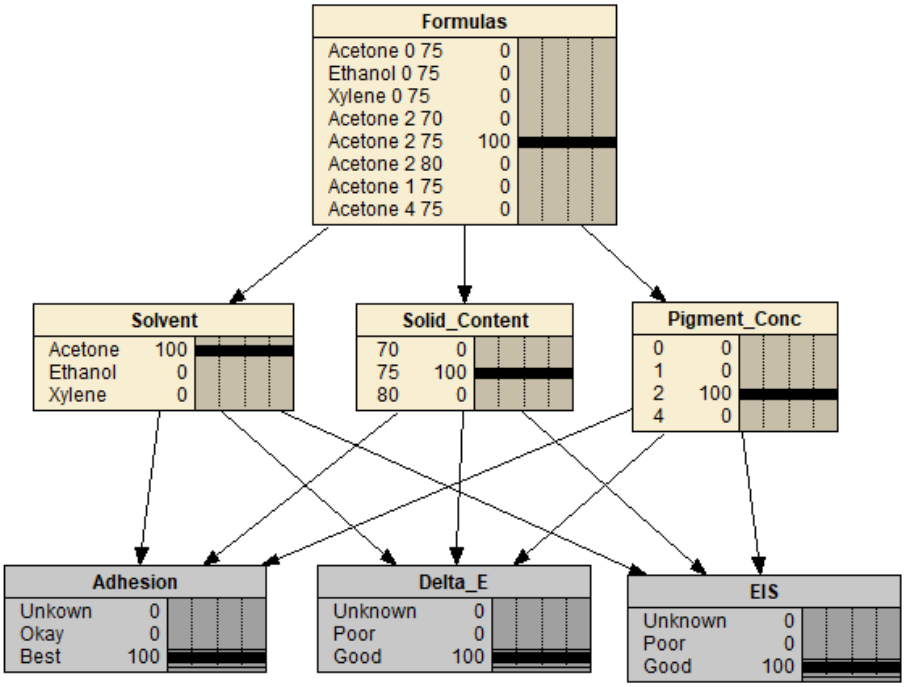


Figure 15

This Netica model demonstrates that the ideal coating uses Acetone as a solvent and has 2% pigment concentration and 75% solids content.

Discussion and Next Steps

Bayesian Networks show great promise as decision making tools for both marketing purposes and experimental formulation. Additional work should be completed to expand the usability of the models in this project. The marketing model can be expanded to model other products with larger data sets. This model can be improved by using a user interface that abstracts the Bayesian Network from the user; by this method, the user would input their own level of importance for each property and the model will recommend the best choices for their needs. This would be useful for in-store or online applications by preventing the customer from having to manually sort through many choices. The Sherwin-Williams expert knowledge model can be improved by inputting data from a formulation study to determine optimal amounts of each pigment. This can then be expanded to implement other paint components to build a tool that can help formulators to estimate the properties of prototype blends without having to mix each test batch. Similarly, development should continue with the epoxy formulation model as more data becomes available. The largest need for improvement is in the ranking method for EIS performance, of which other methods should be evaluated. When larger data sets become available, a critical initial value for impedance should be determined to separate good coatings from bad. After this threshold is determined, coatings with similar initial values can be compared with the gradient method.

References

[1] Ayello, F.; Jain, S.; Sridhar, N; Koch, G.H. (2104) "Quantitative Assessment of Corrosion Probability- A Bayesian Network Approach" *NACE International – Corrosion; Vol. 70. No. 11*.

[2] Dal Ferro, N.; Quinn, C.; Morari, F. (2018). "A Bayesian belief network framework to predict SOC dynamics of alternative management scenarios" *Soil & Tillage Research; 179: 114-124.*

[3] "EIS of Organic Coatings and Paints" *Gamry Instruments*. 4-22-18 https://www.gamry.com/application-notes/EIS/eis-of-organic-coatings-and-paints/

[4] Fu, C.; Deng, S.; Jin, G.; et. al. (2016). "Bayesian network model for identification of pathways by integrating protein interaction with genetic interaction data." *BMC Systems Biology; 11(Suppl 4):81*.

[5] Huang, C.; Darwiche, A. (1996). "Inference in Belief Networks: A Procedural Guide" *International Journal of Approximate Reasoning*; 15:225-263.

[6] Koleske, Joseph V. (2012). "Paint and Coating Testing Manual - 15th Edition of the Gardner-Sward Handbook" *ASTM International*.

[7] Mokrzycki, Wojciech; Tatol, Maciej. (2011). "Color difference Delta E - A survey" *Machine Graphics and Vision*; 20. 383-411.

[8] Murphy, Kevin. (1998) "A Brief Introduction to Graphical Models and Bayesian Networks" *University of British Columbia.* www.cs.ubc.ca/~murphyk/Bayes/bnintro.html

[9] "Netica – Limited (free) Mode" *Norsys Software Corp.* [Computer Software] 2017.

[10] Sutton, Kaylee. *The Sherwin-Williams Company*. [Personal Communication]. 2017.

[11] *The Sherwin-Williams Company*. [In-store marketing materials]. Cuyahoga Falls, OH. 10-9-17.

[12] Xie, G.; Gao, H.; Qian L.; et. al. (2018). "Vehicle Trajectory Prediction by Integrating Physics- and Maneuver-Based Approaches Using Interactive Multiple Models" *IEEE Transactions on Industrial Electronics; 65:7:5999-6008.*

[13] Zhou, Qixin. *The University of Akron*. [Personal Communication]. 2018.